PATENT ABSTRACTS OF JAPAN

(11)Publication number : 05-303391
(43)Date of publication of application : 16.11.1993

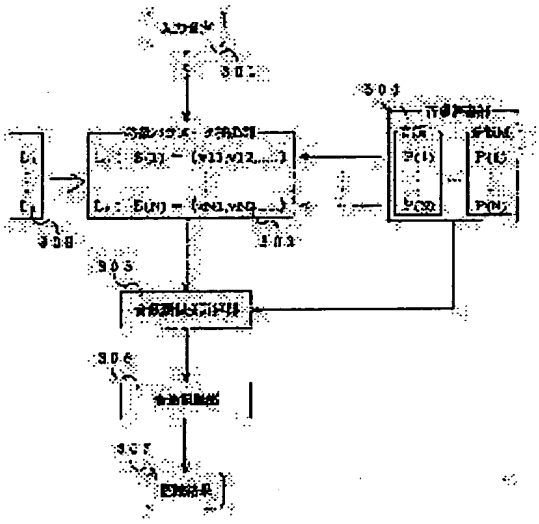
(51)Int.Cl. G10L 3/00
G10L 3/00
G10L 9/00

(21)Application number : 04-106895
(22)Date of filing : 24.04.1992

(71)Applicant : SEIKO EPSON CORP
(72)Inventor : HASEGAWA HIROSHI

(54) SPEECH RECOGNITION DEVICE

(57)Abstract:
PURPOSE: To keep a high recognition rate by respectively computing feature parameter time sequences from plural frame lengths and performing respective phoneme collating.
CONSTITUTION: A feature parameter computing section 303 generates N feature parameter time sequences S1 to SN based on a frame length table 302. 16-th order LPC cepstrum coefficients are computed from 12-th order linear prediction coefficients LPC and are made to 16-dimensional feature vectors. Then, a frame is shifted to compute a feature vector. By repeating these processes, feature parameter time sequences S(j) are obtained. So far as a phoneme model is concerned, N standard patterns are made from feature parameters computed from L1 to LN frame lengths and a phoneme dictionary section 304 is constituted. A phoneme similarly computing section 305 performs a phoneme collation, outputs the phoneme with a presence probability higher than a threshold value to a phoneme recognition section 306, checks presence position.presence probability of phonemes, leaves the larger presence probability ones and outputs them.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平5－303391

(43)公開日　平成５年(1993)11月16日

(51)Int.Cl.[5]	識別記号		FI
G10L 3/00	515	Z 8842-5H	
	531	D 8842-5H	
		E 8842-5H	
9/00	301	A 8842-5H	

審査請求　未請求　請求項の数３　（全７頁）

(21)出願番号　特願平4－106895

(22)出願日　平成４年(1992)４月24日

(71)出願人　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号

(72)発明者　長谷川　浩
長野県諏訪市大和３丁目３番５号セイコーエプソン株式会社内

(74)代理人　弁理士　鈴木　喜三郎　（外１名）

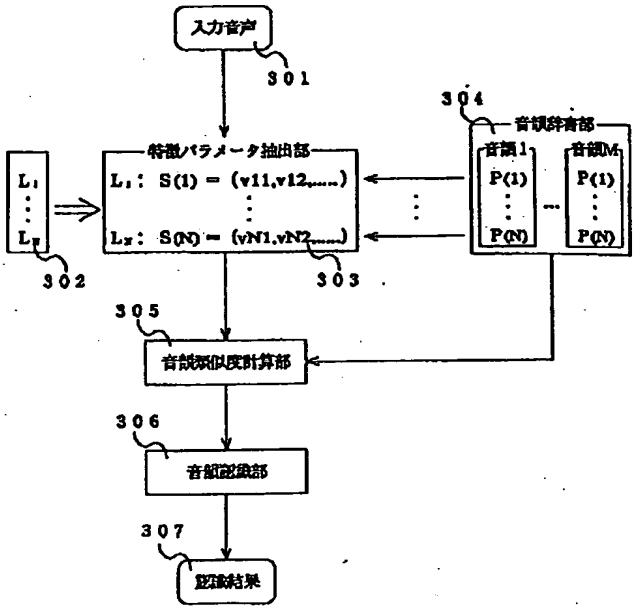
(54)【発明の名称】音声認識装置

(57)【要約】

【目的】　入力音声から特徴パラメータを計算する際、各音韻の継続時間長の違いを反映させることにより、音韻識別率の向上をはかる。

【構成】　特徴パラメータを計算するための単位時間（フレーム）を複数個用意する、あるいは各音韻毎に用意し、各フレーム長毎に特徴パラメータ時系列を計算し、そのそれぞれに対して音韻照合を行い、最適なものを選ぶ。

【効果】　各音韻にとっての最適なフレーム長を用いることによって、音韻認識率が向上する。また、複数のフレーム長によって特徴パラメータを計算することによって、入力音声の時間的変動に対しても誤認識が少ない。



1

【特許請求の範囲】

【請求項１】　入力された音声データから、単位時間（１フレーム）分のデータごとに特徴パラメータを計算し、その特徴パラメータ時系列と、音韻辞書部における各音韻の標準パターンとの類似度を認識部で求め、類似度の高い音韻を認識結果とする音声認識装置において、あらかじめ特徴パラメータを計算するための単位時間長（フレーム長）を複数個（Ｎ個）備え、それぞれの音韻毎に、フレーム長L_1により作成した標準パターンＰ（１）から、フレーム長L_iにより作成した標準パターンＰ（Ｎ）までのＮ個ずつの標準パターンを持つ音韻辞書部を備え、

入力された未知音声から、前記のＮ個のフレーム長L_1～L_iを用いてＮ個の特徴パラメータ時系列Ｓ（１）～Ｓ（Ｎ）を計算する特徴パラメータ計算部と、

各音韻毎に、Ｎ個すべての特徴時系列に対して、時系列Ｓ（ｉ）と前記標準パターンＰ（ｉ）の類似度を計算し、得られたＮ個の類似度の最大値をその音韻の類似度として出力する類似度計算部と、

前記類似度計算部で出力された各音韻ごとの類似度を比較し、最も類似度の高い音韻を認識結果として出力する音韻認識部と、

を備えることを特徴とする音声認識装置。

【請求項２】　入力された音声データから、単位時間（１フレーム）分のデータごとに特徴パラメータを計算し、その特徴パラメータ時系列と、音韻辞書部における音韻の標準パターンとの類似度を音韻認識部で求め、類似度の高い音韻を認識結果とし、前記音韻認識結果から音韻系列を生成し、前記音韻系列と語彙辞書の内容の類似度を求めて単語や文節を認識する音声認識装置において、

あらかじめ各音韻ごとに異なる長さの単位時間長（フレーム長）のデータから計算される特徴パラメータによって作成された音韻辞書部を備え、各音韻毎に最適なフレーム長およびフレームをシフトさせる時間幅を用いて特徴パラメータを計算する特徴パラメータ計算部を、各音韻毎に備えることを特徴とする音声認識装置。

【請求項３】　音韻を、その継続時間長により複数のグループに分け、個々のグループ毎に異なる長さの単位時間長（フレーム長）のデータから計算される特徴パラメータによって作成された音韻辞書部を備え、各音韻グループ毎に異なるフレーム長およびフレームをシフトさせる時間幅を用いて特徴パラメータを計算する特徴パラメータ計算部を備えることを特徴とする音声認識装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は音韻単位の認識に基づく音声認識装置に関するものである。

【０００２】

【従来の技術】現在考案されている音声認識装置は、そのほとんどが音声を特徴量の時系列に変換し、その時系列をあらかじめもっている標準パターンの時系列と比較して認識を行うというものである（図１、図２）。特徴量を計算する場合、通常数ミリ秒から数十ミリ秒を単位時間（これをフレームという）とし、１フレームの時間内では特徴量すなわち音声の波の構造は定常状態にあると近似して、ＬＰＣケプストラム等の特徴パラメータを計算する。そしてフレームをある時間だけずらして（これをフレームシフトという）、ふたたび特徴パラメータを計算する。これを繰り返すことによって特徴パラメータの時系列が得られ、これを標準パターンと比較、類似度を計算することによって認識が行われる。（たとえば特開昭61-238099）しかし音声信号は本来動的な性質をもっており、刻々とその状態は変化している。そのため１フレームの時間長（これをフレーム長という）が長すぎると、例えば／ｐ／／ｔ／／ｋ／といった短い継続時間の子音の特徴をとらえることが難しくなる。逆にフレーム長が短いと、データ数が少なくなるため特徴パラメータの推定の精度が悪くなったり、音声にとって重要な波長の長い波の成分が見えにくくなったりする、という問題が生じてくる。

【０００３】この問題を解決するために、フレーム毎の特徴パラメータの差分ベクトルを特徴量に加え、場合によって差分ベクトルの方を重視する（特開平03-145167）等の工夫がなされているが、同じフレーム長で母音・子音の特徴パラメータをとらえているため、子音に関してはフレーム長が長すぎるため特徴量が隠され、母音に関してはフレーム長が短かすぎるためパラメータの短時間のゆらぎ等による誤認識がおこる可能性があった。これらの問題は特徴量の差分ベクトルを用いるだけでは解決できなかった。

【０００４】

【発明が解決しようとする課題】本発明の課題は、音韻認識率を大きく左右する特徴パラメータの計算方法を改善し、各音韻の認識率を向上することである。

【０００５】

【課題を解決するための手段】上記課題を解決するため、本発明の音声認識装置は、入力された音声データから、単位時間（１フレーム）分のデータごとに特徴パラメータを計算し、その特徴パラメータ時系列と、音韻辞書部における各音韻の標準パターンとの類似度を認識部で求め、類似度の高い音韻を認識結果とする音声認識装置において、あらかじめ特徴パラメータを計算するための単位時間長（フレーム長）を複数個（Ｎ個）備え、それぞれの音韻毎に、フレーム長L_1により作成した標準パターンＰ（１）から、フレーム長L_iにより作成した標準パターンＰ（Ｎ）までのＮ個ずつの標準パターンを持つ音韻辞書部を備え、入力された未知音声から、前記のＮ個のフレーム長L_1～L_iを用いてＮ個の特徴パラメータ時系列Ｓ（１）～Ｓ（Ｎ）を計算する特徴パラメー

タ計算部と、各音韻毎に、Ｎ個すべての特徴時系列に対して、時系列Ｓ（ｉ）と前記標準パターンＰ（ｉ）の類似度を計算し、得られたＮ個の類似度の最大値をその音韻の類似度として出力する類似度計算部と、前記類似度計算部で出力された各音韻ごとの類似度を比較し、最も類似度の高い音韻を認識結果として出力する音韻認識部と、を備えることを特徴とする。

【０００６】また、入力された音声データから、単位時間（１フレーム）分のデータごとに特徴パラメータを計算し、その特徴パラメータ時系列と、音韻辞書部における音韻の標準パターンとの類似度を音韻認識部で求め、類似度の高い音韻を認識結果とし、前記音韻認識結果から音韻系列を生成し、前記音韻系列と語彙辞書の内容の類似度を求めて単語や文節を認識する音声認識装置において、あらかじめ各音韻ごとに異なる長さの単位時間長（フレーム長）のデータから計算される特徴パラメータによって作成された音韻辞書部を備え、各音韻毎に最適なフレーム長およびフレームをシフトさせる時間幅を用いて特徴パラメータを計算する特徴パラメータ計算部を、各音韻毎に備えることを特徴とする。

【０００７】また音韻を、その継続時間長により複数のグループに分け、個々のグループ毎に異なる長さの単位時間長（フレーム長）のデータから計算される特徴パラメータによって作成された音韻辞書部を備え、各音韻グループ毎に異なるフレーム長およびフレームをシフトさせる時間幅を用いて特徴パラメータを計算する特徴パラメータ計算部を備えることを特徴とする。

【０００８】

【作用】本発明は以上の構成を有するので、子音に対しては短いフレーム長で、また母音に対しては長いフレーム長で計算された特徴パラメータ用いて認識を行うことが可能となる。

【０００９】

【実施例】

（実施例１）以下本発明を実施例に基づいて詳述する。

【００１０】音韻認識をおこなう場合、それぞれの音韻の平均継続時間長が問題となる。音韻の特徴は、おおきく「語頭（前の音韻の影響をうける部分）」「語中（その音韻固有の部分）」「語尾（後の音韻の影響をうける部分）」の３つに分けられる。文献（音響学会講論集1-2-14(1988-03)）によると、特に短い／ｔ／／ｒ／などの子音は語頭・語中・語尾の平均継続時間長は１５ミリ秒程度しかないのに対し、母音の方はそれぞれ１００ミリ秒を越える平均時間長をもつ。このように継続時間長に大きなばらつきがある様々な音素を認識するため、本発明ではあらかじめ特徴パラメータを計算するためのフレーム長を複数用意することで対処する。

【００１１】図３は本発明の構成を示す図である。サンプリング周波数２０ｋＨｚで１６ビットで量子化された入力音声（３０１）から、特徴パラメータ計算部（３０３）において、特徴パラメータを計算する。特徴パラメータ計算部（３０３）は、あらかじめ用意されているＮ個のフレーム長L_i(i=1...N)が記述されているフレーム長テーブル（３０２）に従って、Ｎ個の特徴パラメータ時系列Ｓ(1)～Ｓ(N)を生成する。この手順を説明したのが図２である。２０１はデジタル化された入力音声信号である。まずこの入力信号の先頭からフレーム長L_i（２０３）分のデータに注目し、この中のデータを定常状態にあるとみなして特徴パラメータを計算する。本実施例においては１２次の線形予測係数ＬＰＣから１６次のＬＰＣケプストラム係数を計算して１６次元の特徴ベクトル（２０２）とした。次にフレームをシフト（２０４）させ、同様に特徴ベクトルを計算する。この操作を入力信号のおわりまでくりかえすことによって、フレーム長L_iを用いた計算した特徴パラメータ時系列Ｓ(i)が得られる。これを全てのフレーム長に関して同様に求める。その結果Ｎ個のパラメータ時系列Ｓ(1)からＳ(N)が得られる。本実施例においてはＮ＝５とし、L_i=3.2, L_i=6.4, L_i=12.8, L_i=25.6, L_i=51.2（いずれもミリ秒）とした。

【００１２】一方あらかじめ個々の音韻モデルに関して、L_iからL_iのフレーム長から計算した特徴パラメータを用いて、Ｎ個の標準パターンを作成しておく。これは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用い、それぞれのフレーム長毎に計算した特徴パラメータ時系列を、個別の隠れマルコフモデル（ＨＭＭ）Ｐ(1)～Ｐ(N)を用意してトレーニングすることによって作成した。こうして得られた音韻数Ｍ×モデル数ＮのＨＭＭモデルにより、音韻辞書部（３０４）を構成した。

【００１３】音韻類似度計算部（３０５）においては、まず各音韻毎に特徴パラメータ時系列Ｓ(1)は標準パターンＰ(1)を用い、Ｓ(2)にはＰ(2)を用い、以下同様にＳ(N)にはＰ(N)を用いて音韻照合を行う。そして、あらかじめ定めたしきい値を上回る存在確率をもつもののみを、音韻認識部（３０６）に出力する。

【００１４】音韻認識部（３０６）では、音韻類似度計算部（３０５）から出力されたすべての音韻の存在位置・存在確率を調べ、存在位置が重なっているものに関しては存在確率の大きなもののみを残す。こうして得られた音韻列を認識結果（３０７）として出力する。

【００１５】本発明により、フレーム長を固定した場合の音韻認識率と比較して、認識率の向上が認められた。

【００１６】（実施例２）図４は本発明の構成を示す図である。４０２は、各音韻を認識する場合最も適当なフレーム長およびフレームシフトの時間長を記述した表である。この表の値は以下のような実験によりあらかじめ求めておく。

【００１７】ここでは音韻／ａ／に関して説明する。音声データベースから音韻／ａ／の発話区間を切り出し、

これを様々なフレーム長・フレームシフト時間長（N個）から計算した特徴量パラメータで、N個の標準パターンを作成する。ここではHMM音韻モデルをそれぞれの標準パターン毎に作成した。そののち、それぞれのHMM音韻モデルを用いて、音韻／ａ／の認識率および／ａ／以外の音韻を／ａ／以外の音韻を／ａ／として認識した誤認識率を調べ、音韻／ａ／の識別率が最も高いものを音韻／ａ／の最適フレーム長・フレームシフトとした。以下全ての音韻に関して同様の実験を行い、音韻最適フレーム長テーブル（４０２）を作成した。

【００１８】このようにして作成された最適フレーム長テーブル（４０２）を用いて、サンプリング周波数２０ｋＨｚで１６ビットで量子化された入力音声（４０１）から、特徴パラメータ計算部（４０３）において、各音韻毎に特徴パラメータを計算する。一般に最適フレーム長・シフトは音韻毎に異なるため、／ａ／の認識のために作成された特徴パラメータ時系列からは、音韻／ａ／だけが照合の対象となる。音韻認識部（４０５）では、音韻／ａ／のために計算された特徴パラメータ時系列から音韻／ａ／を照合し、同様に音韻／ｉ／のための特徴パラメータ時系列から音韻／ｉ／を照合し、以下同様に全ての音韻を、それぞれ別々に計算された特徴パラメータ時系列から照合する。こうして照合された音韻が、認識結果（４０６）として出力される。

【００１９】本発明により、従来方法と比較して演算時間をほとんど増加させることなく、音韻認識率を向上することができた。

【００２０】（実施例３）また、豊富なデータから厳密な実験を行えば、最適なフレーム長・シフトは各音韻毎に異なるが、一般には音韻を、その継続時間長に応じて３ないし４つのグループに分けて、それぞれのグループ毎に共通のフレーム長・シフトを用いれば、期待した効果が十分得られることが多い。実施例３として、音韻を「母音」「半母音・拗音」「破裂音」「その他の子音」の４グループにわけ、それぞれについてはフレーム長・シフトを共通の値とした。これにより認識率をほとんど保ったまま、演算時間を若干短縮することができた。

【００２１】

【発明の効果】本発明は複数のフレーム長から特徴パラメータ時系列をそれぞれ計算し、そのそれぞれから音韻照合を行う。音声認識装置に対する入力音声は発声速度、話者の違い等により同じ音韻でもその時間的特徴は様々に変化する。従来はこういった入力に対してまず同一フレーム長を用いて特徴量を計算し、その後の処理で入力音声のばらつきを吸収するという方法がとられている。これに対し本発明では特徴量を計算する段階で様々なフレーム長を用い、その結果最適なものを音韻ごとに選択して用いる、という手法を用いるものである。このため従来の方法と比較して入力音声の時間的特徴がばらついても、高い認識率を維持することが可能となる。

【００２２】さらに、あらかじめ各音韻にとって最適なフレーム長・フレームシフト時間長を実験によって求め、その値を用いて音韻認識をすることによって、従来方法と比較してより正確な音韻認識を行うことが可能である。

【００２３】また、最適フレーム長・フレームシフト時間長が類似した音韻をグループ化することによって、本手法を高速化することが可能である。

【００２４】また、本発明は音声認識のデータの前処理の部分に関するものであるため、隠れマルコフモデル（ＨＭＭ）等の公知のマッチング技術と組み合わせることによって、従来よりも高い認識率が得られる。

【図面の簡単な説明】

【図１】　従来の音声認識の処理を説明する図。

【図２】　特徴パラメータ抽出部の処理を説明する図。

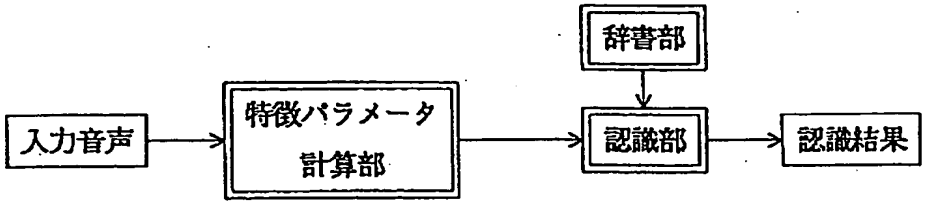
【図３】　本発明の処理を説明する図。

【図４】　本発明の処理を説明する図。

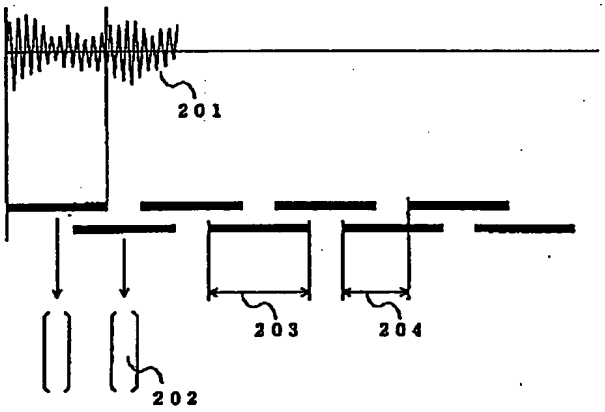
【符号の説明】

２０１　入力音声
２０２　特徴パラメータベクトル
２０３　フレーム長
２０４　フレームシフト
３０１　入力音声
３０２　フレーム長の表
３０３　特徴パラメータ計算部
３０４　音韻辞書部
３０５　音韻類似度計算部
３０６　音韻認識部
３０７　認識結果
４０１　入力音声
４０２　音韻最適フレーム長テーブル
４０３　特徴パラメータ計算部
４０４　音韻辞書部
４０５　音韻認識部
４０６　認識結果

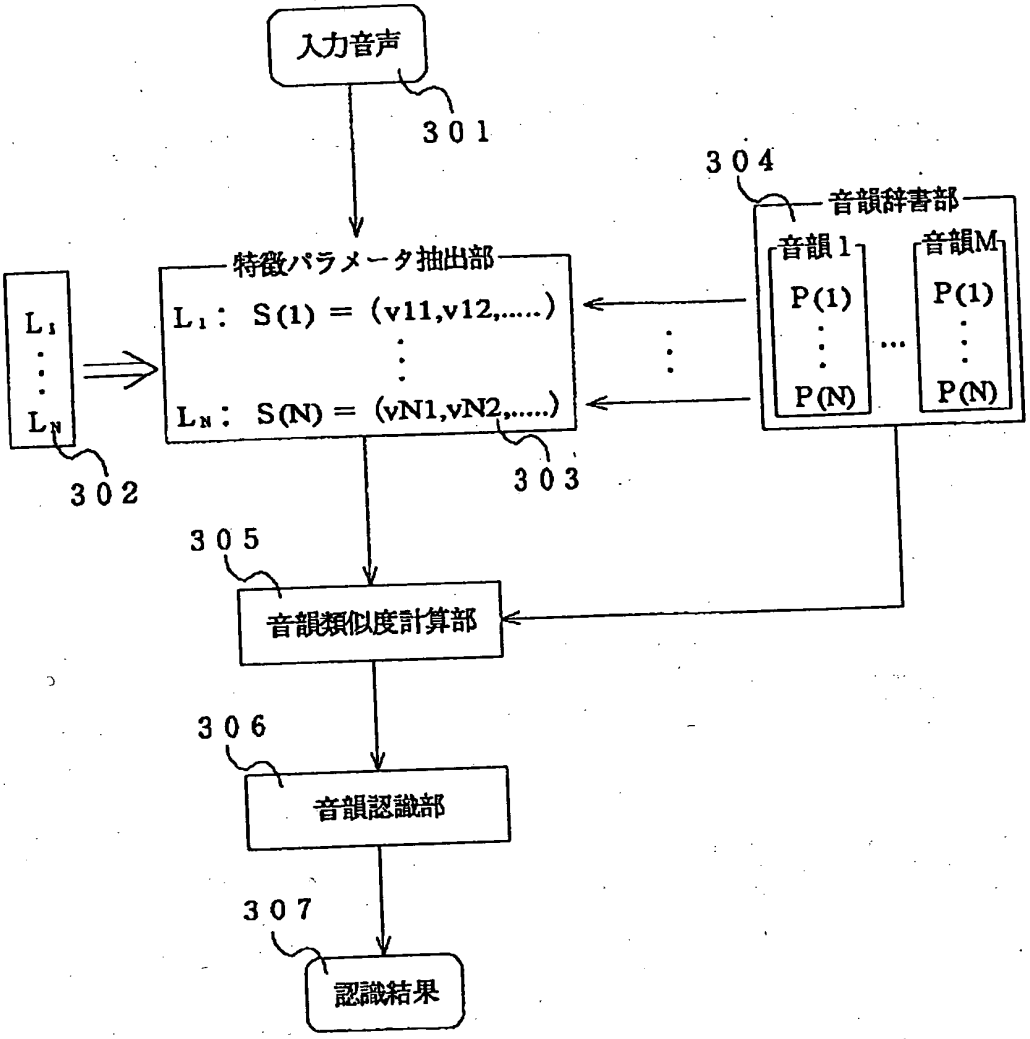
【図１】



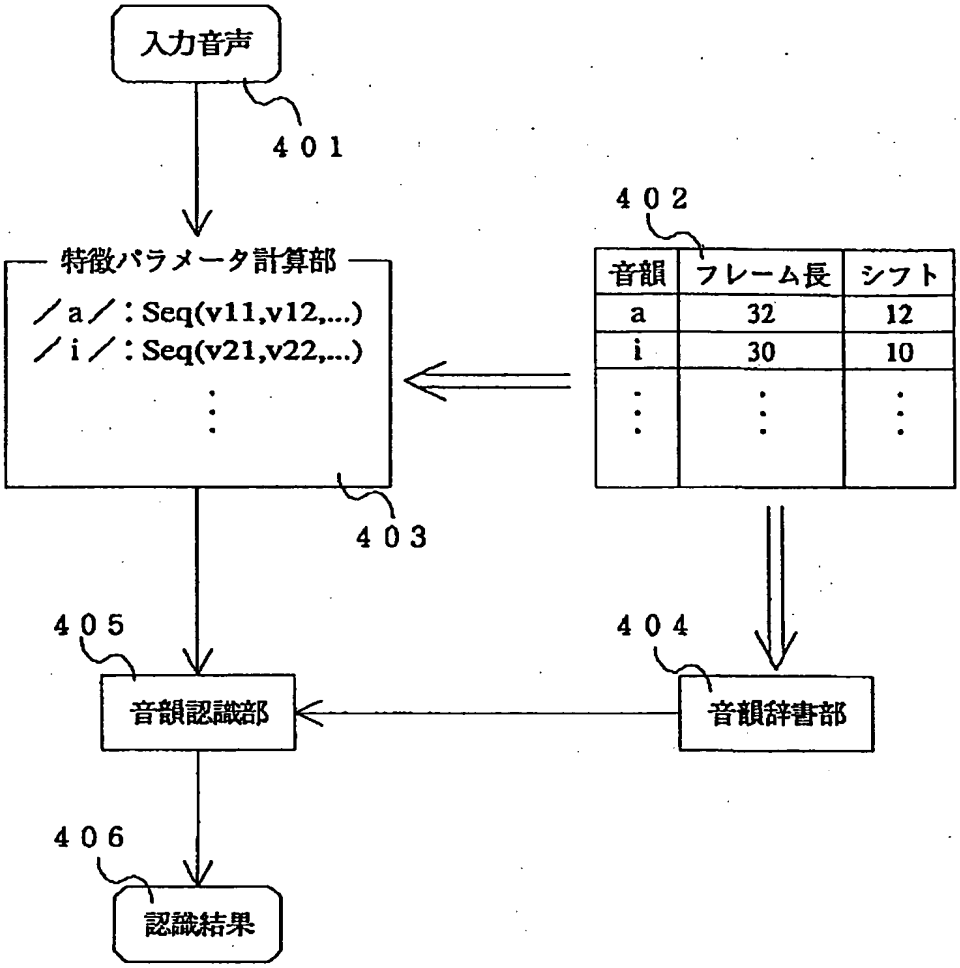
【図２】



【図３】


入力音声
３０１
３０４
音韻辞書部
音韻１
音韻M
P(1)
P(N)
特徴パラメータ抽出部
L₁ : S(1) = (v11,v12,.....)
LN : S(N) = (vN1,vN2,.....)
L₁
LN
３０２
３０３
３０５
音韻類似度計算部
３０６
音韻認識部
３０７
認識結果

【図４】


入力音声
401
402
音韻
フレーム長
シフト
a
32
12
i
30
10
特徴パラメータ計算部
／a／：Seq(v11,v12,...)
／i／：Seq(v21,v22,...)
403
405
404
音韻認識部
音韻辞書部
406
認識結果